Speech Emotion Recognition and Perception of Music

Mélanie Fernández Pradier

Prof. Dr.-Ing. Bin Yang

Supervisors: Prof. Dr.-Ing. Bin Yang
Dipl.-Ing. Fabian Schmieder

January 27, 2011

Motivation

Speech Emotion Recognition and Perception of Music

Emotion Recognition from Speech

- Speech $\sim$ two-channel
  - linguistic
  - paralinguistic
- Several Applications
  - support ASR
  - diagnoses
  - speech synthesis
  - entertainment

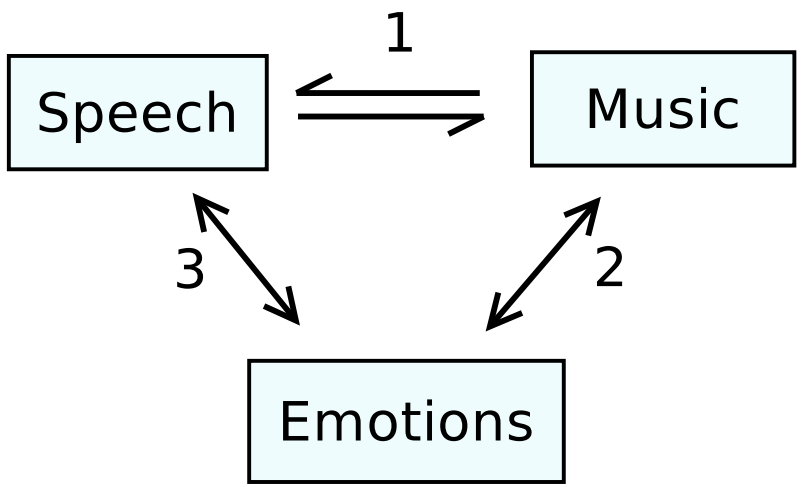
Music Perception

- "language of emotion"
- treatment of affective disorders
- treatment of speech disorders
- same origin of music and speech

Aim of the thesis

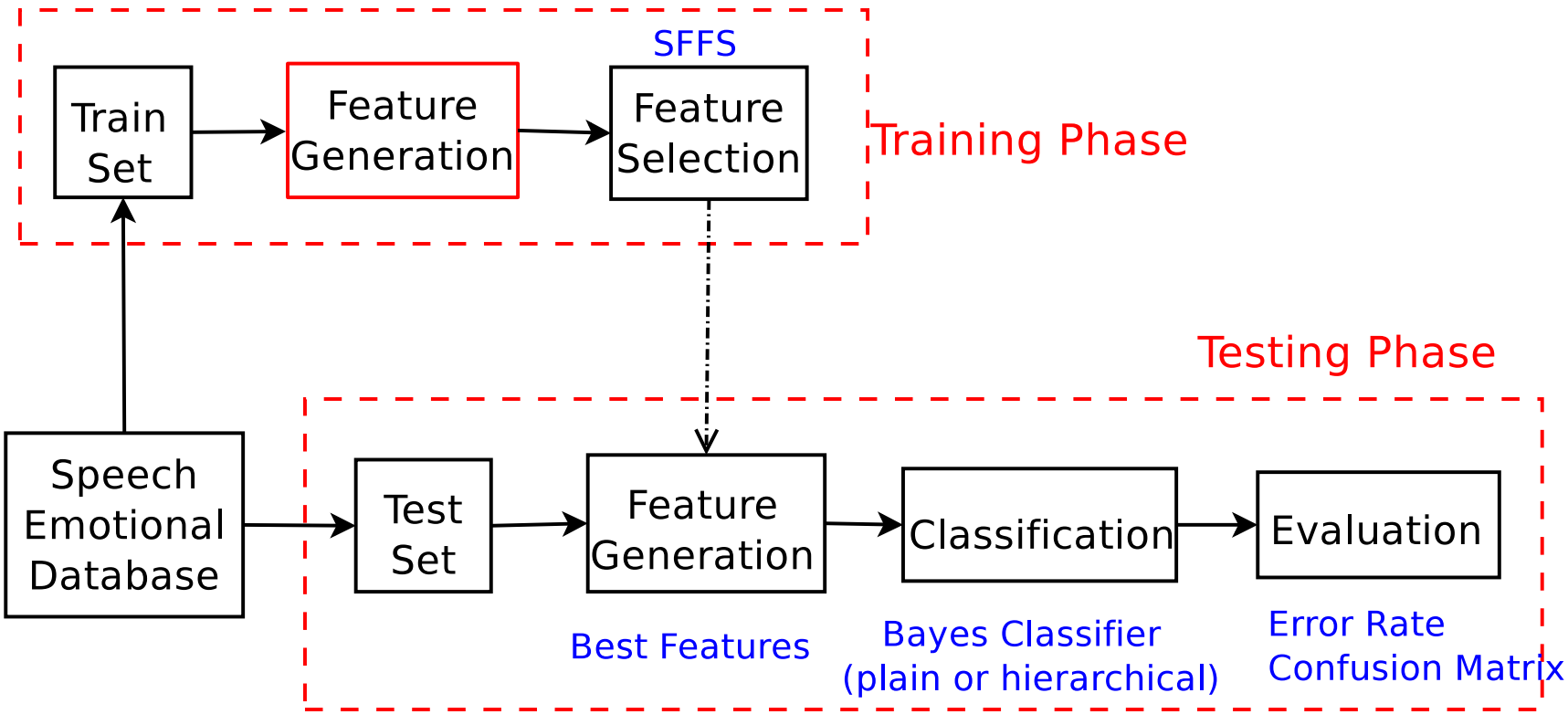
Apply Music Theory to Speech Emotion Recognition

Investigate Speech and Music similarities to derive universal features for Emotions

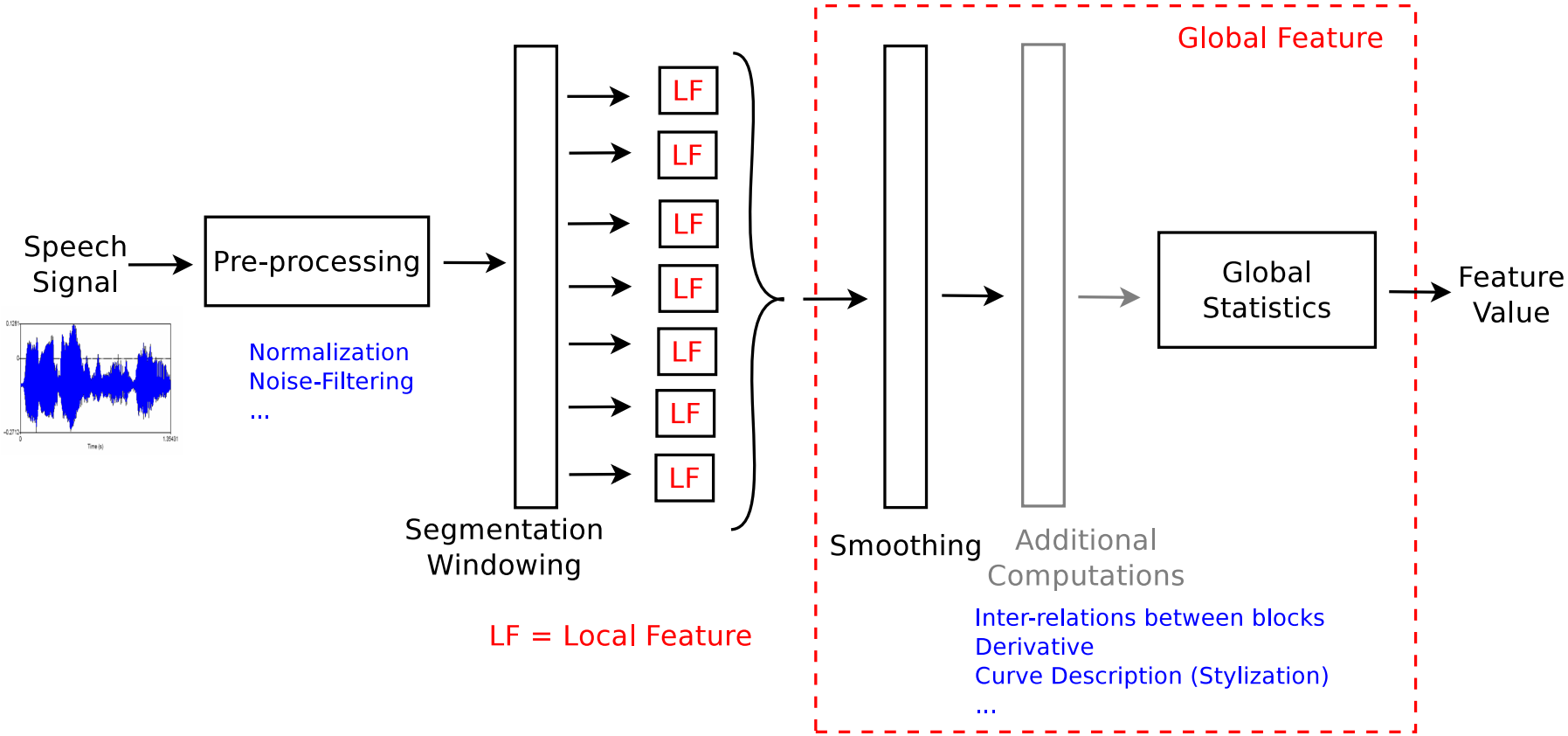


1. What is the link between music and speech?
2. How are emotions transmitted through music?
3. Can we apply musical knowledge to speech processing?

Pattern Recognition



Feature Generation



Basic Features Description

Local Features

- ZCR
- MFCC
- Energy
  total + bands
- Pitch
- Voiced-unvoiced
- VAD

$$ZCR = \frac{1}{2} \cdot \sum_{n=1}^{N} |\mathrm{sgn}(x_n) - \mathrm{sgn}(x_{n+1})|$$

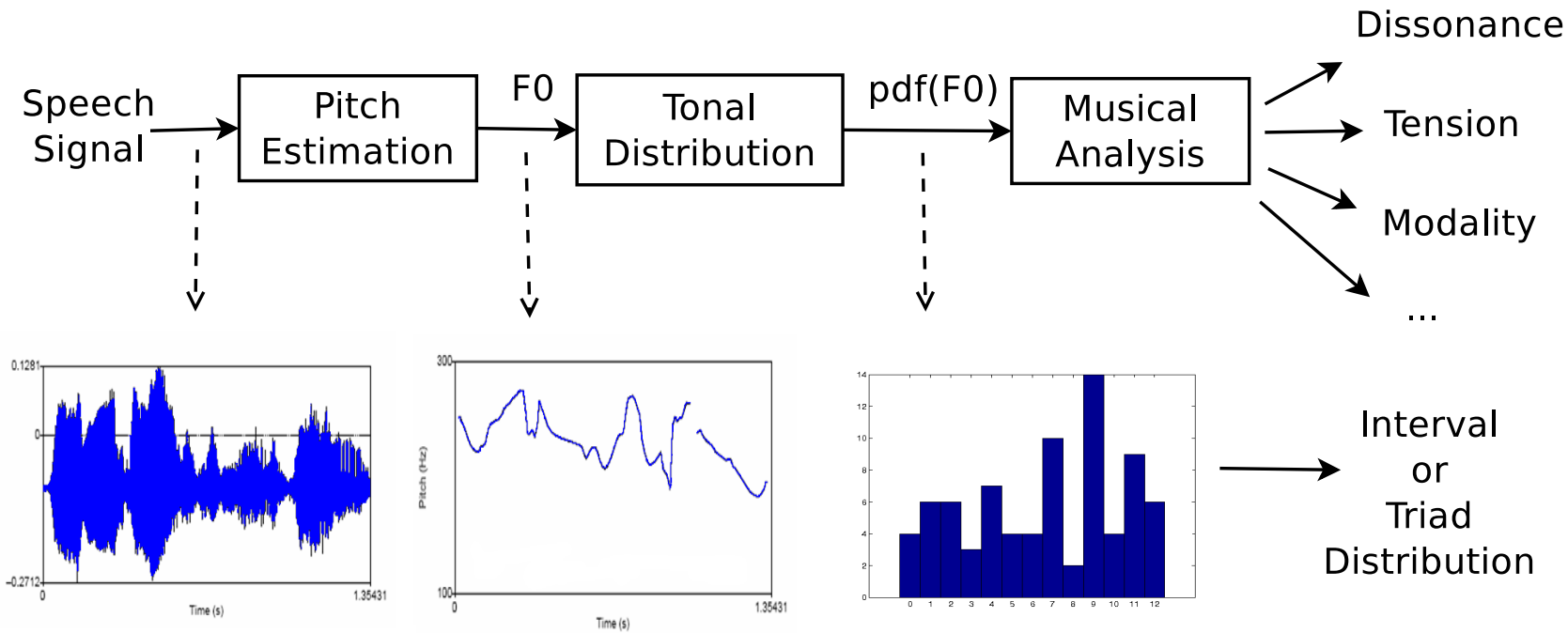
$$Cepstrum = \left| FFT\left\{ \log\left(|FFT\{\underline{x}\}|^2 \right) \right\} \right|^2$$

$$Energy = \sum_{n=1}^{N} x_n \cdot x_n^{\star}$$

Global Features

- Global statistics: min, mean, max, median, std, iqr...
  directly, 1st or 2nd derivative
- Energy and pitch plateaux
- Combination with logical features

Interval and Triad Features



Interval Features

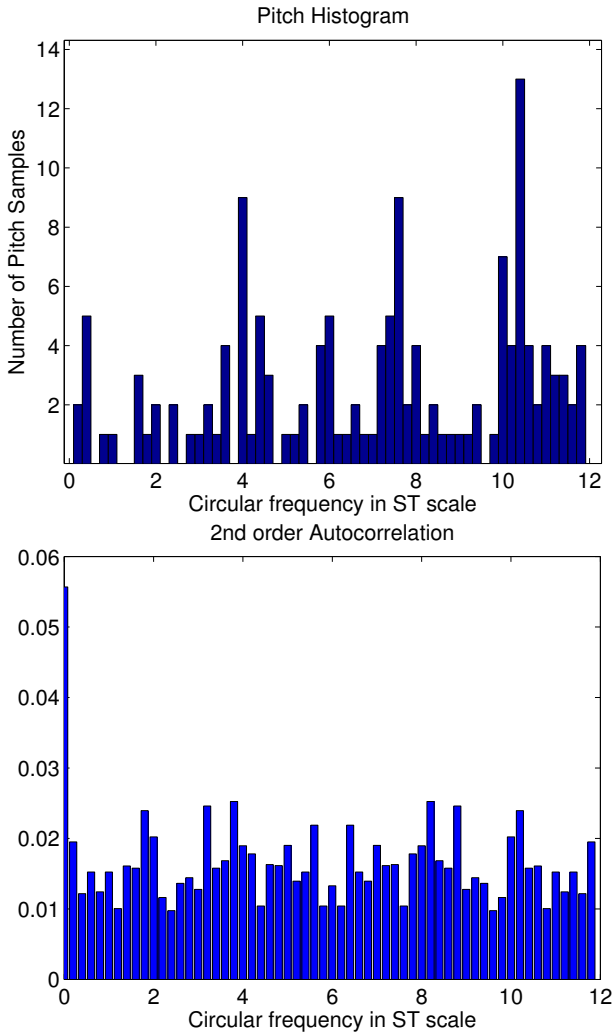
Autocorrelation of the circular pitch density function

$$\int_0^L p_o\left(\mathrm{mod}_L\left(s+\lambda\right)\right) p_o\left(\lambda\right) d\lambda$$

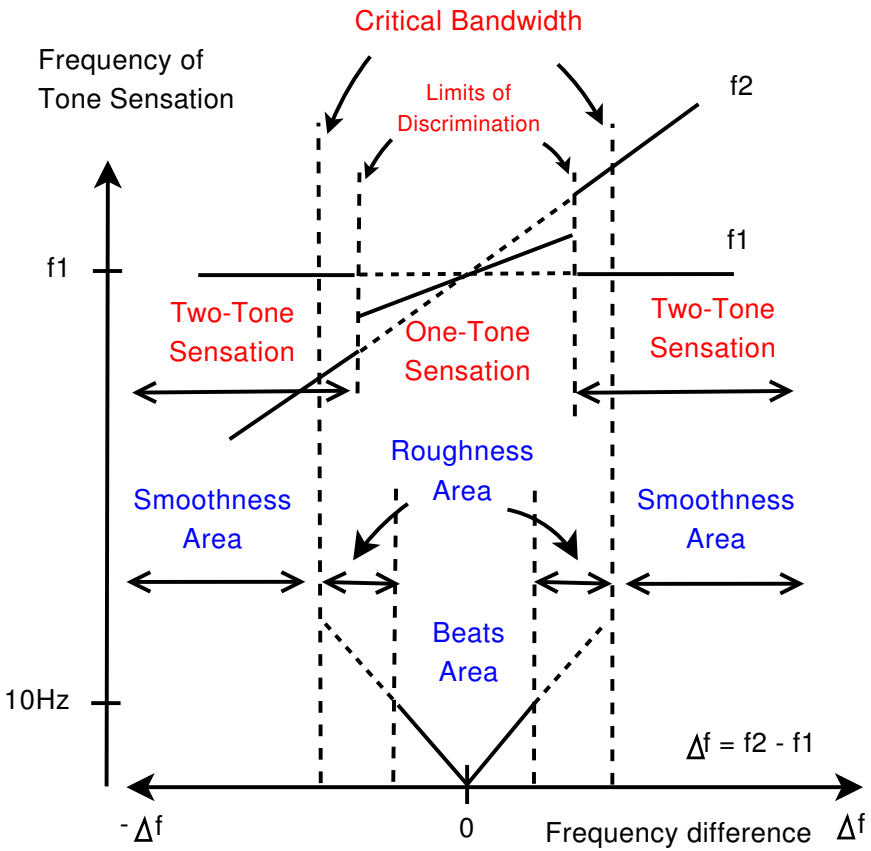
Intervalic dissonance

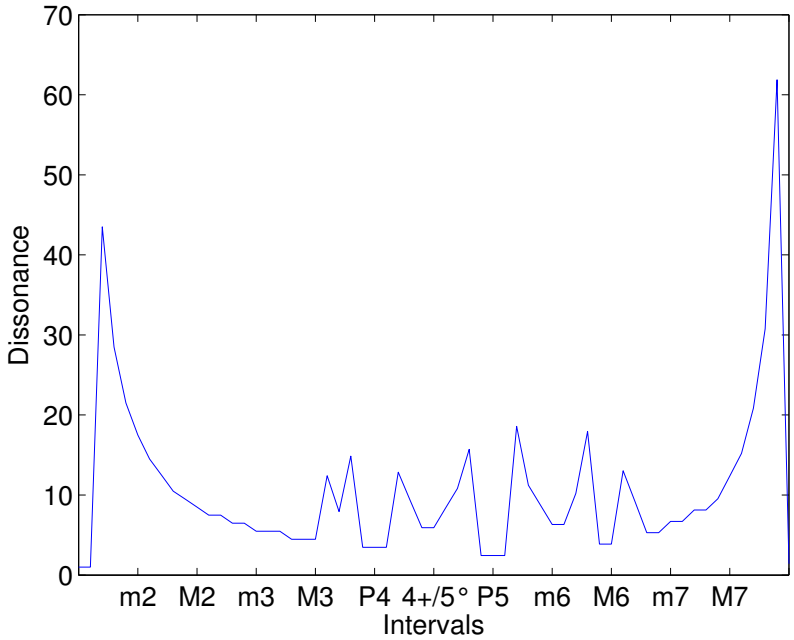
$$DIS = \int_0^L d\left(s\right) r_o\left(s\right) ds$$

$$where \;\; d\left(s\right) \simeq \sqrt{N\left(s\right) D\left(s\right)}$$



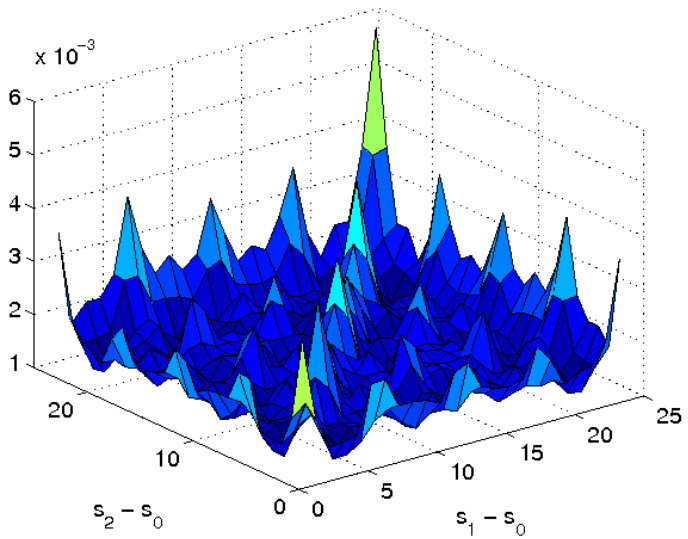
Interval Dissonance



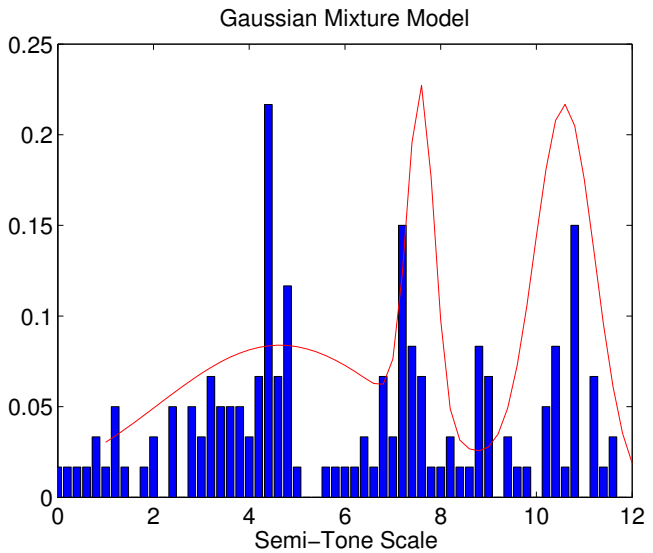


Triad Features

1. Direct computation
2. Extraction of "dominant pitches"

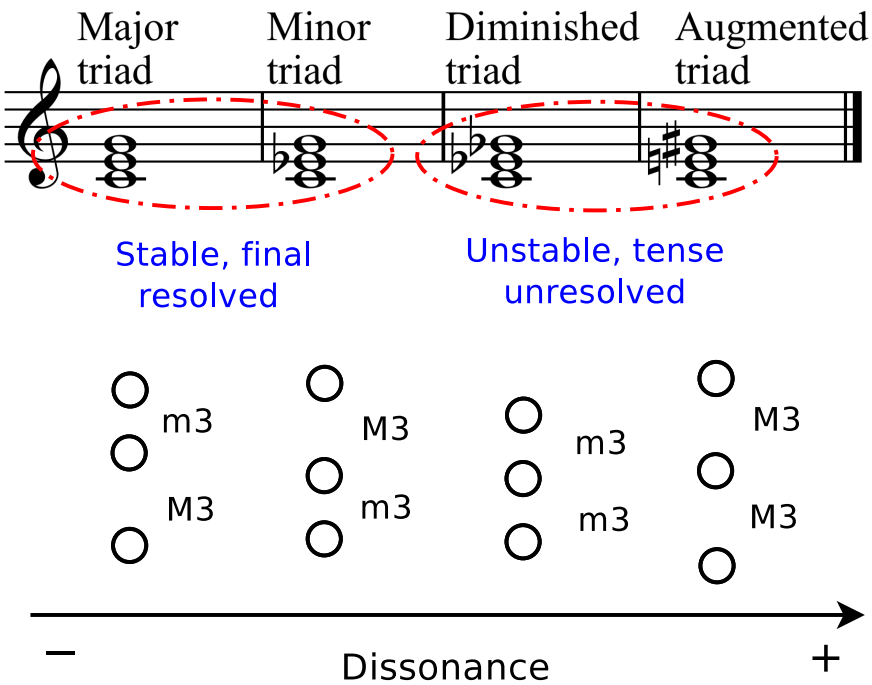


Autocorrelation Triad Features



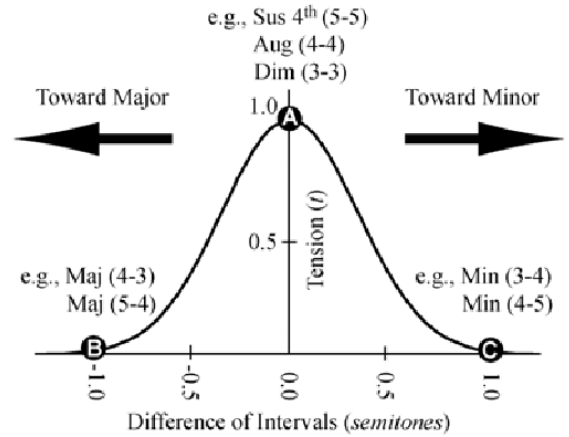
Gaussian Triad Features

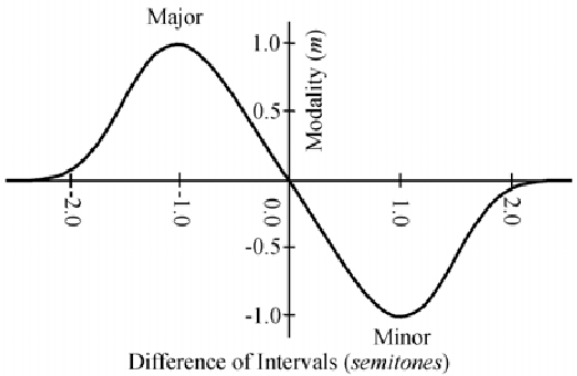
Tension and Modality



m3 = minor third (3 ST)

M3 = major third (4 ST)





Loudness, Timbre and Rhythm

Intensity Features

$$I(k) = \sum_{n=0}^{N/2} |\mathrm{FFT}_k(n)|$$

$$D_i(k) = \frac{1}{I(k)} \sum_{n=L_i}^{H_i} |\mathrm{FFT}_k(n)|$$

where k refers to the frame

Timbre Features

$$\mathrm{FFT}_k \equiv \{x_{k1} \dots x_{kN}\}$$

$$\rightsquigarrow \text{sorted} \equiv \left\{x'_{k1} \dots x'_{kN}\right\}$$

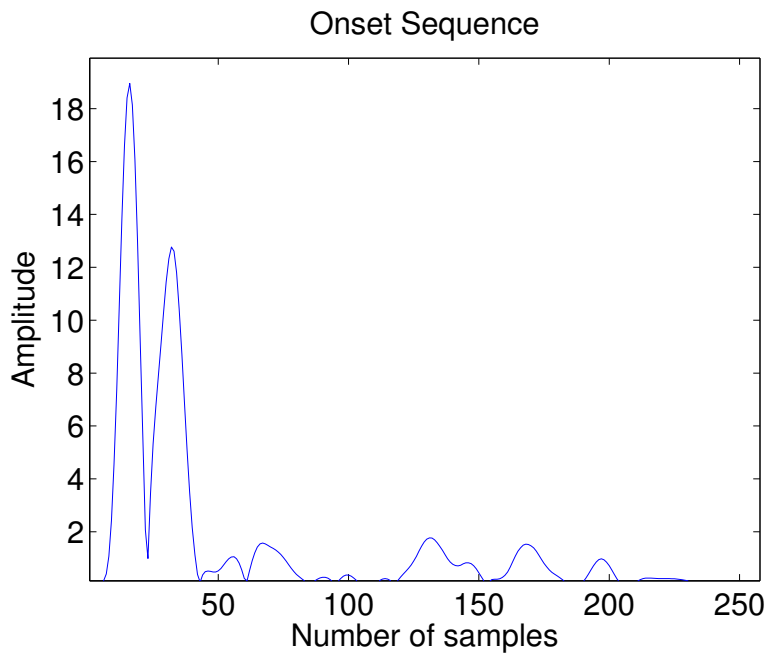
$$Peak(k) = \log\left\{\frac{1}{\alpha N} \sum_{i=1}^{\alpha N} x'_{ki}\right\}$$

$$Valley(k) = \log\left\{\frac{1}{\alpha N} \sum_{i=1}^{\alpha N} x'_{k(N-i+1)}\right\}$$

Loudness, Timbre and Rhythm

Rhythm Features

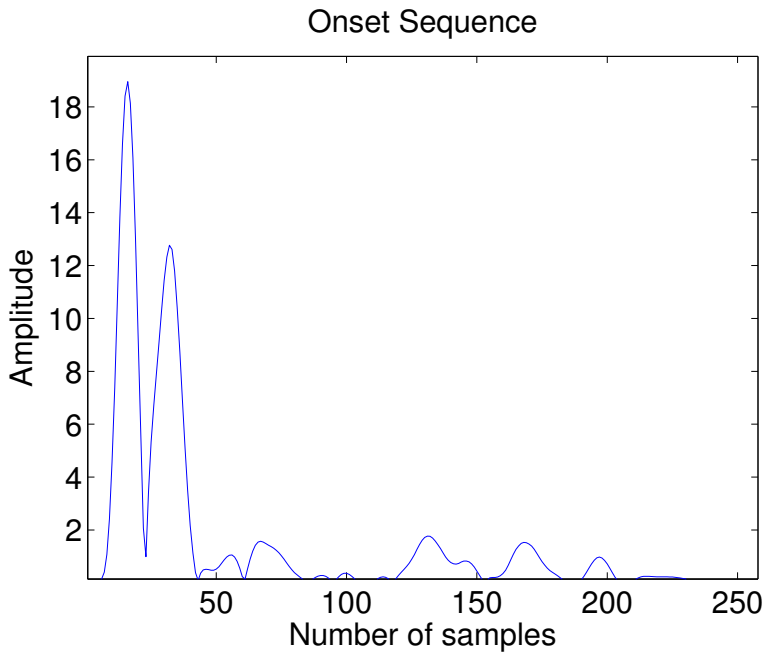
1. Compute FFT
2. Extract amplitude envelope
   $A_i(n) = FFT_i(n) \otimes h_w(n)$
3. Apply Canny operator
   $O_i(n) = A_i(n) \otimes C(n)$
   $C(n) = \frac{n}{\sigma^2} e^{-\frac{n^2}{2\sigma^2}}$
   We obtain the onset sequence
   $O_i(n)$



Loudness, Timbre and Rhythm

Rhythm Features

Strength	Average value of the peaks
Regularity	Average value of peaks in the autocorrelation
Speed	Ratio of number of peaks and time duration



Perceptual Model of Intonation

Perceptual principles

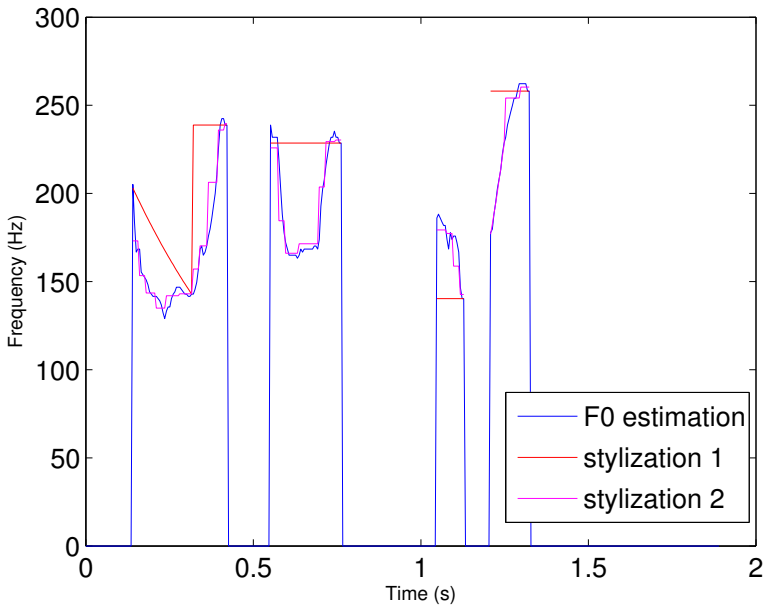
1. Segmentation Effect
2. Glissando Threshold: minimum amount of frequency change

$$g_{th} = 0.16/T^2 \ [\mathrm{ST/s^2}]$$

3. Differential Glissando Threshold: minimum difference in slope

$$dg_{th} = a_2 - a_1 = 20 \ [\mathrm{ST/s}]$$

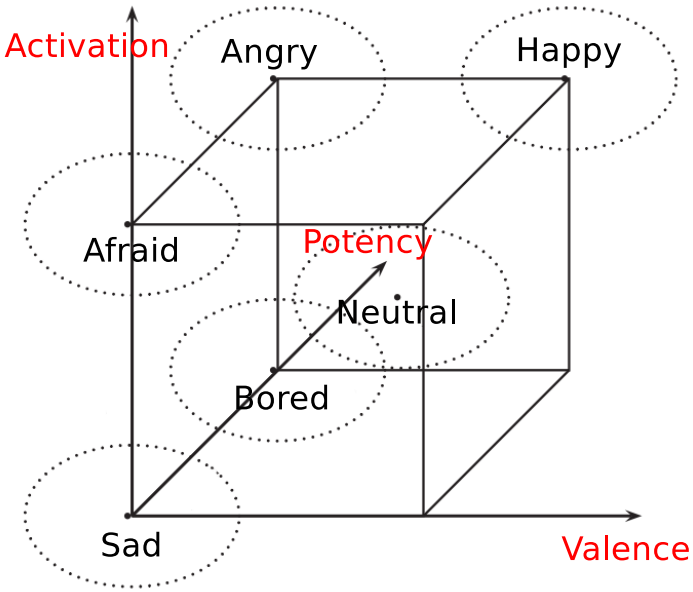
4. Short-term integration in time



Database - Labels - Features

Database: emoDB (TUB)

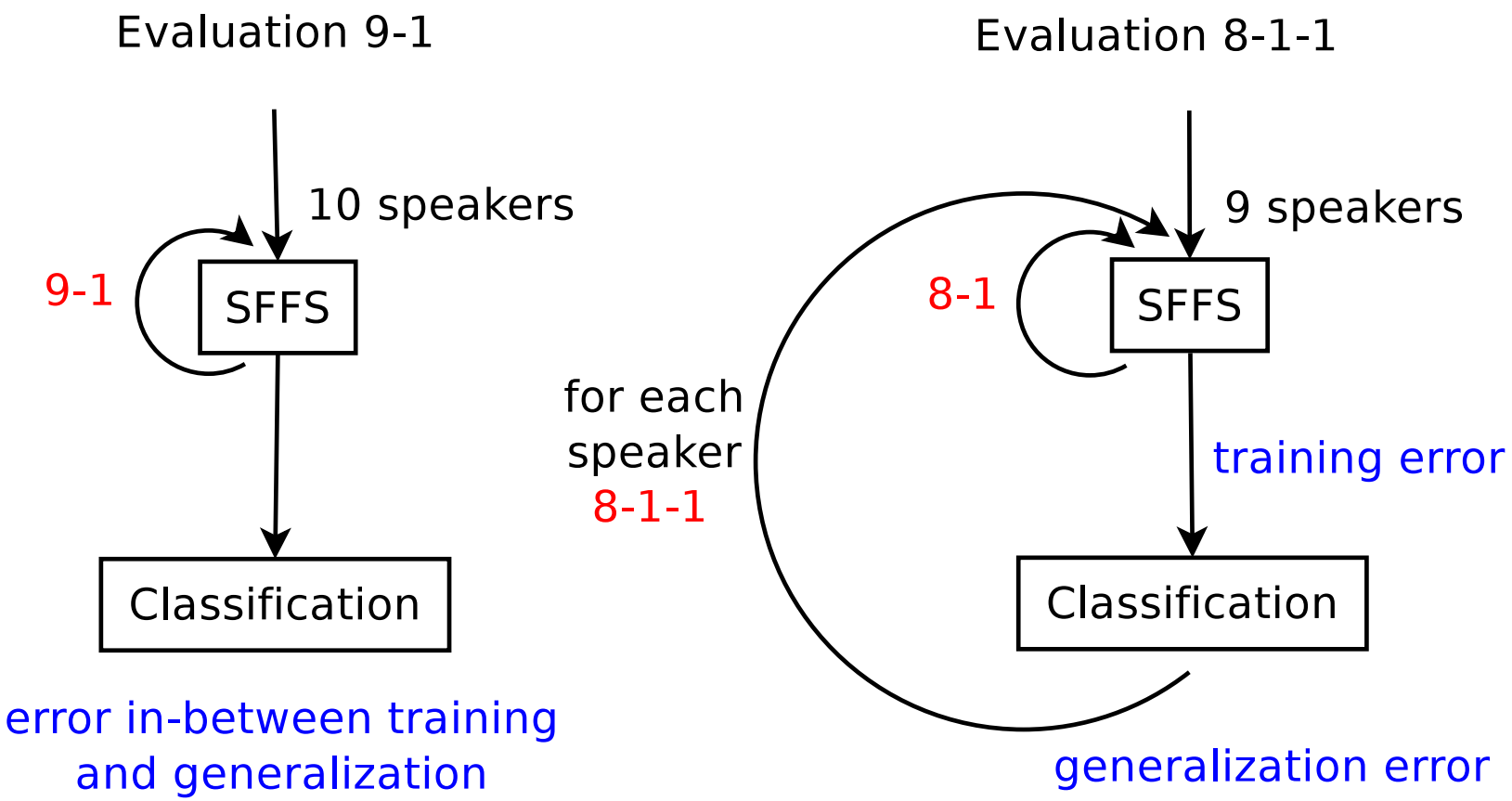
- 10 speakers
- 708 files
- 6 emotions



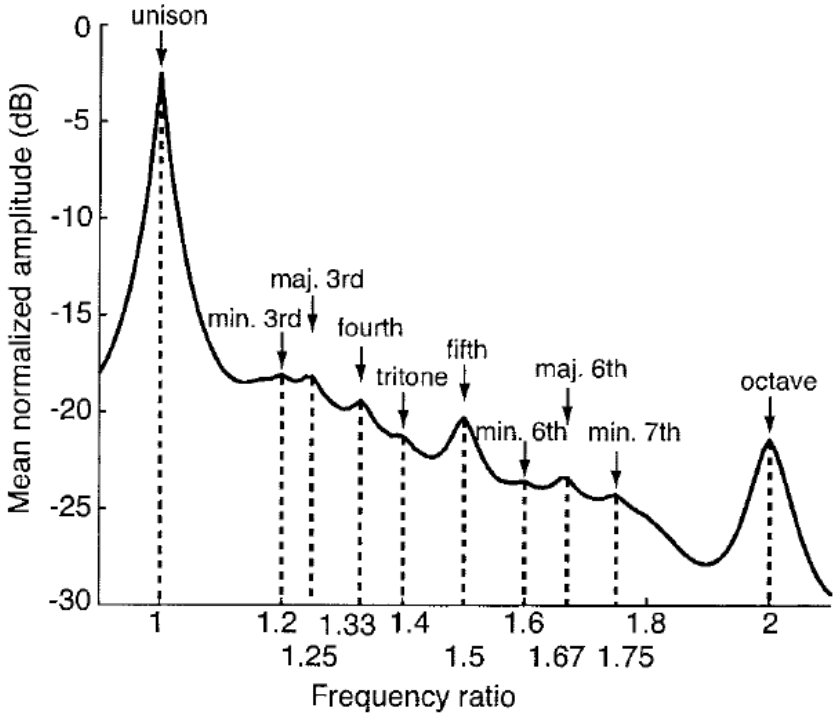
BASIC	SET
duration	16
MFCC	91
ZCR	13
harmony	3
energy	58
pitch	33
Total	214

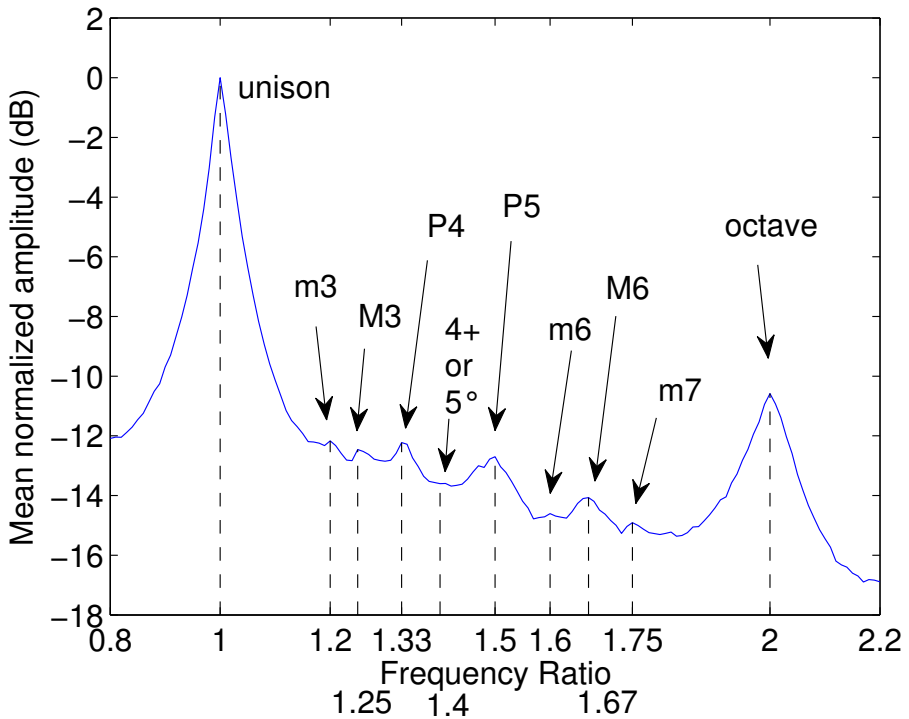
MUSICAL	SET
interval	31
autocorr. triad	4
gaussian triad	10
intensity	63
rhythm	15
Total	123

Strategies for evaluation 9-1 Vs 8-1-1

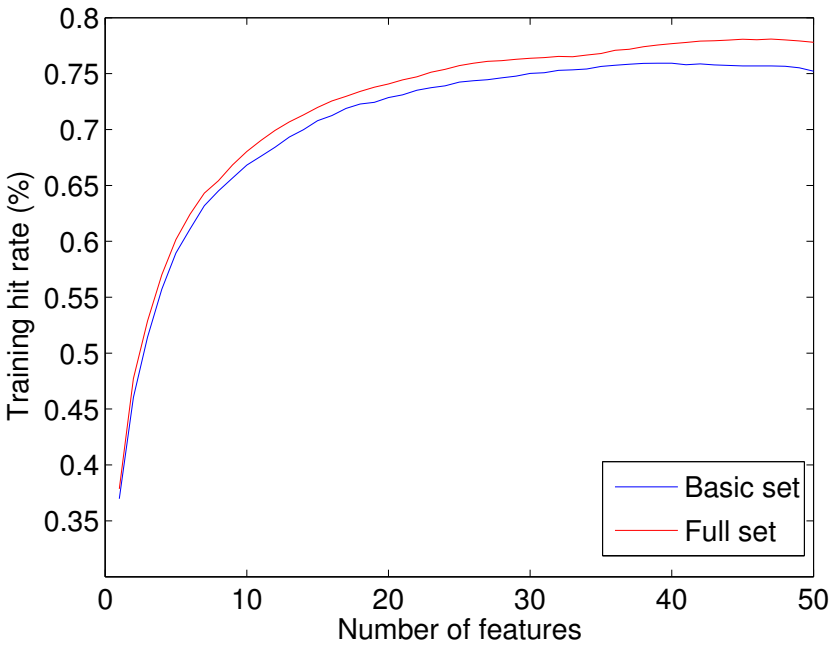


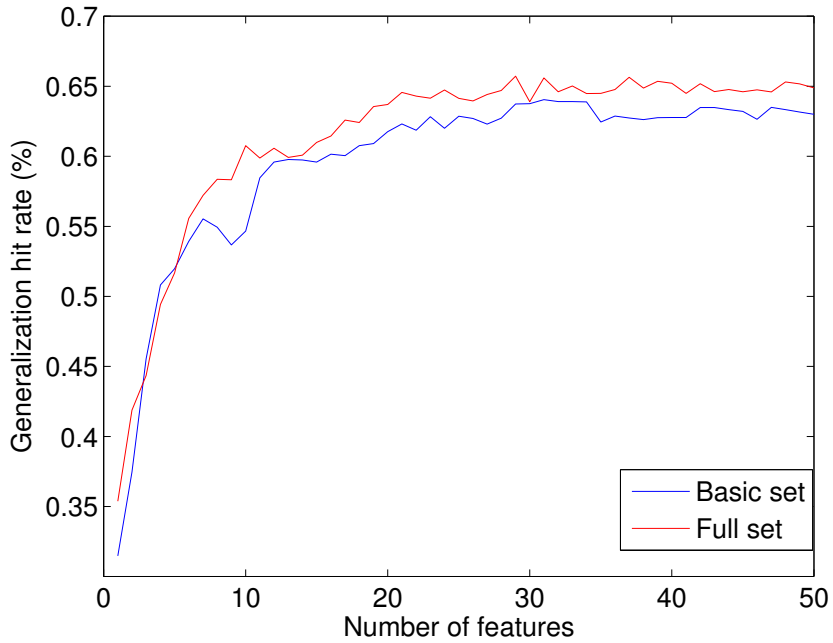
Musical Universals



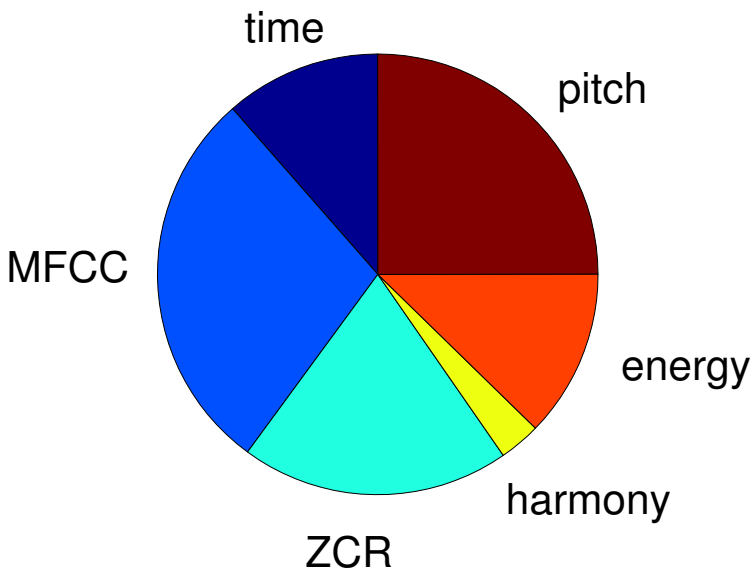


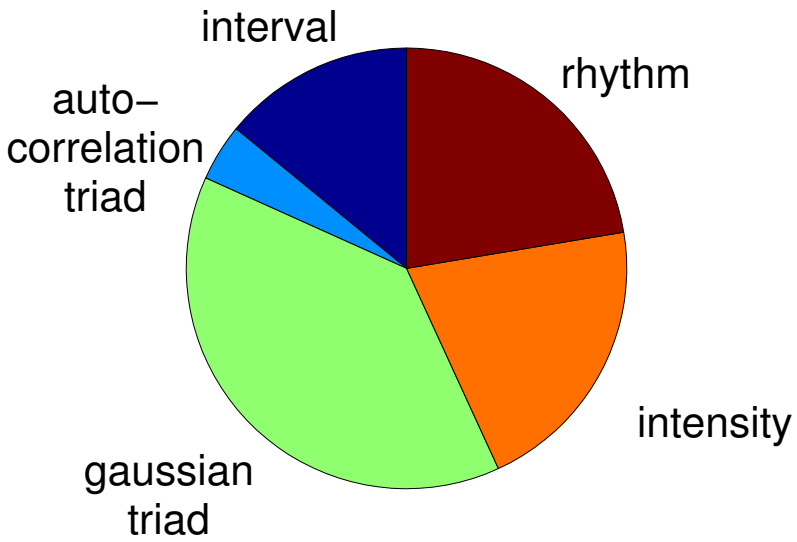
Plain bayes classifier - Evaluation 8-1-1



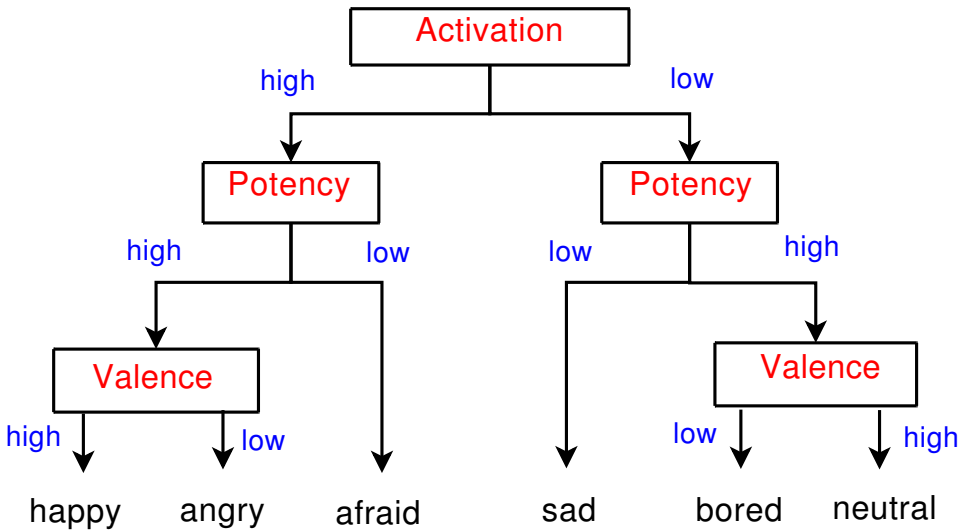


Nature of selected features



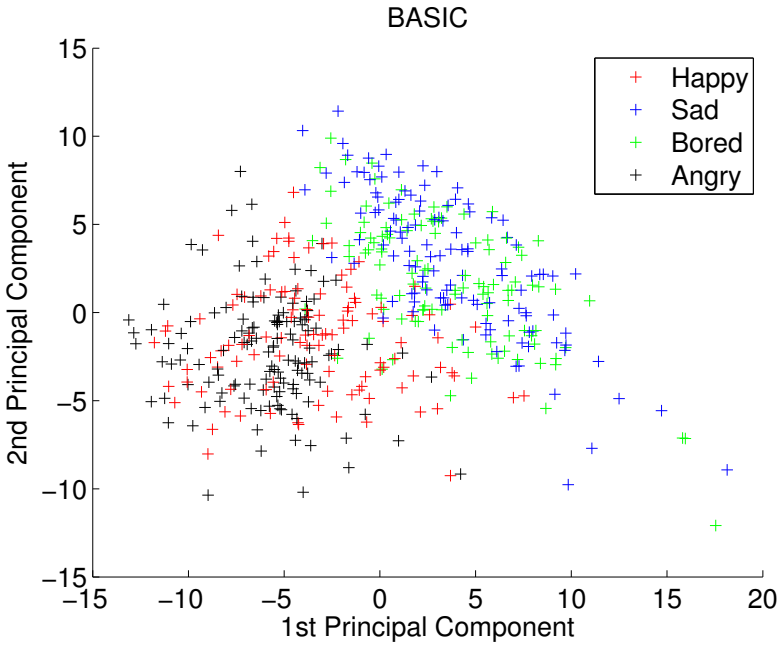


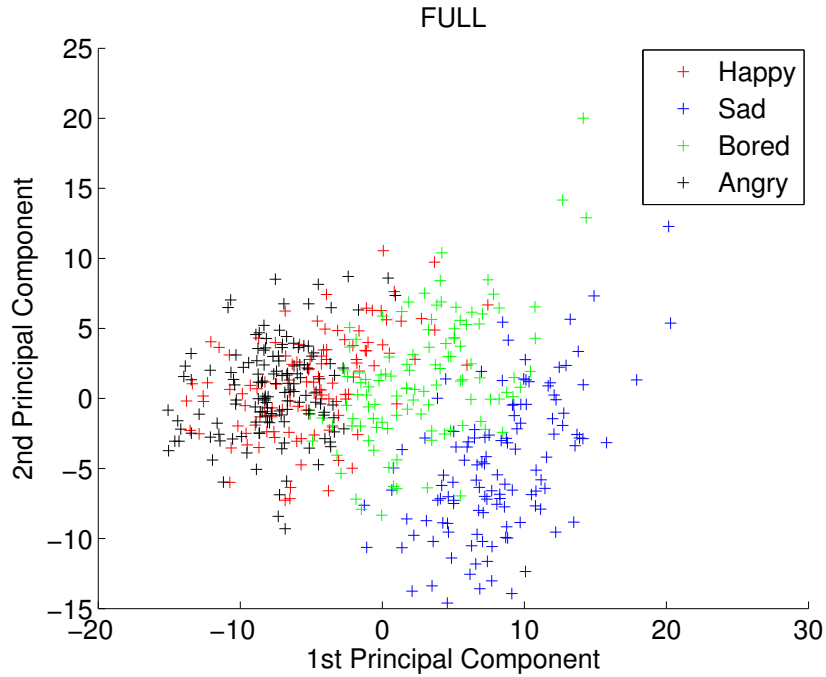
Comparison plain Vs hierarchical bayes classifier



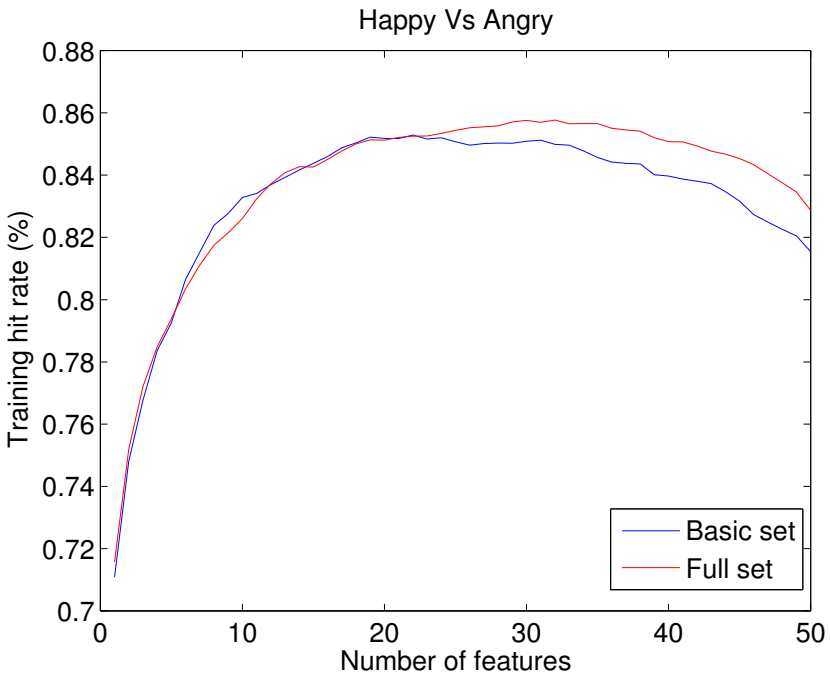
	plain Bayes	hierarchical Bayes
Basic	76.12	84.22
Basic+ Interval+Triad	80.61	85.04

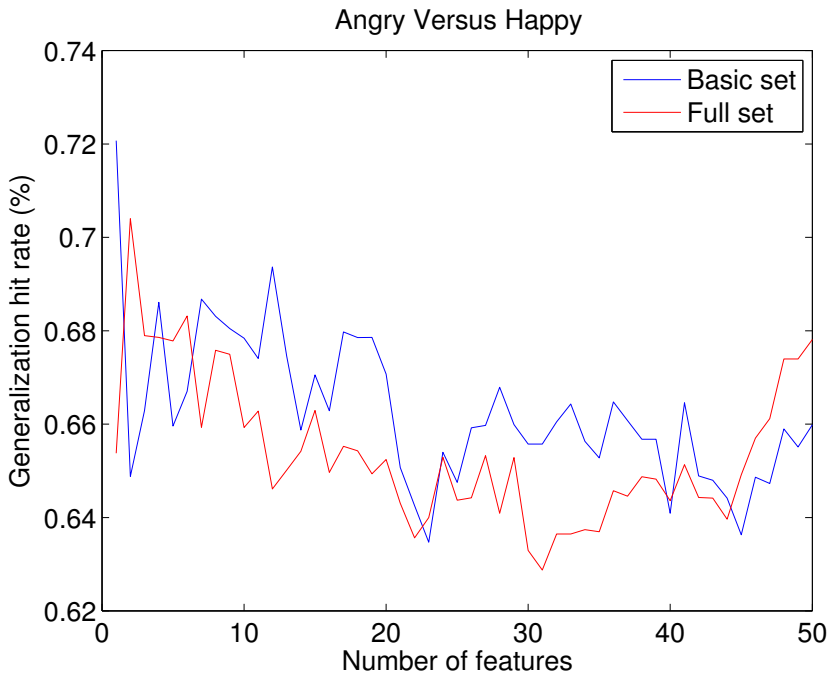
Multi-dimensional Scaling



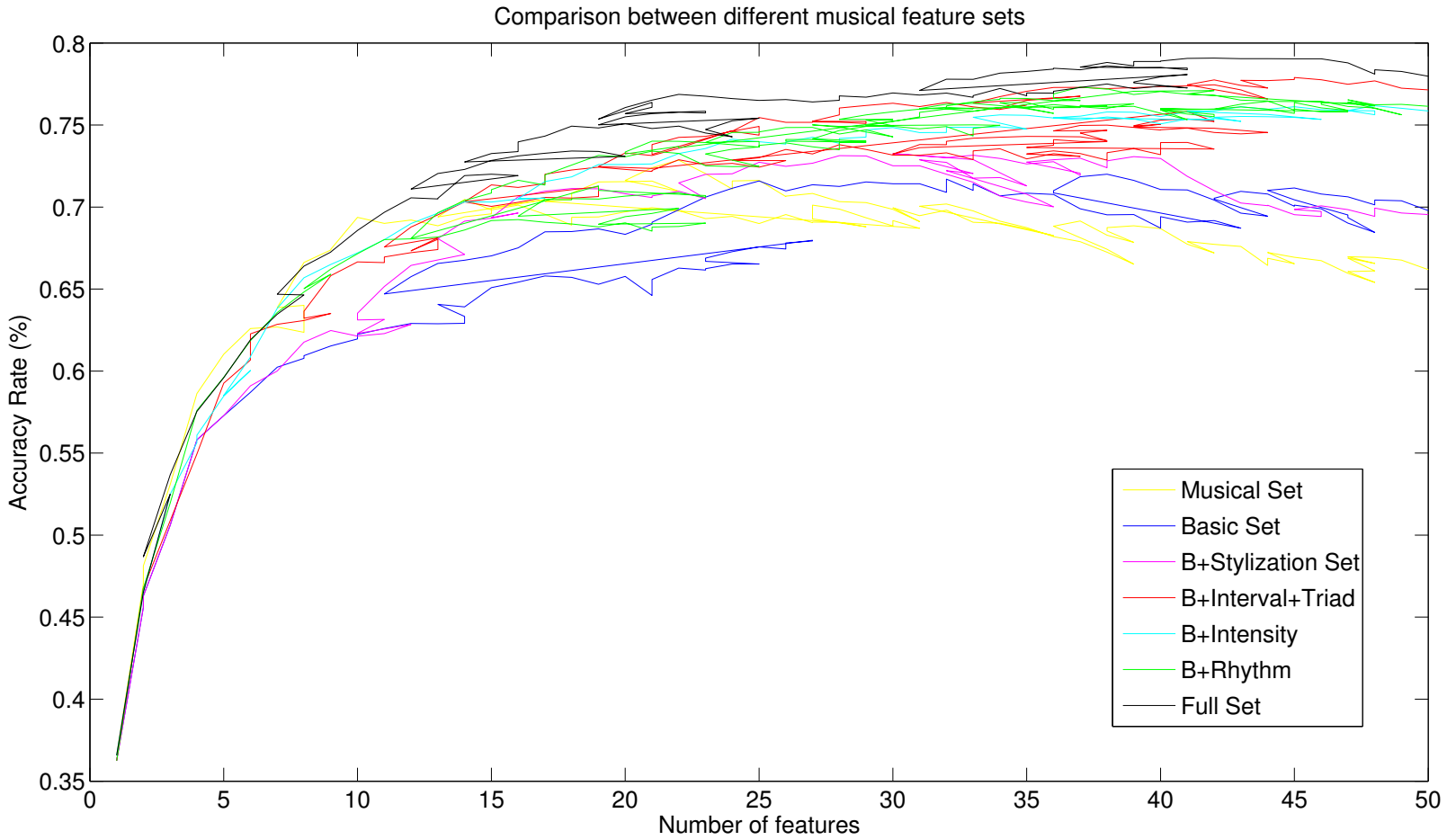


Happy Vs Angry - Evaluation 8-1-1





Final Comparison of Musical Features



Conclusion

Summary

1. Literature review about speech, music and emotions
2. Theoretical background on psychoacoustics
3. Re-implementation of the basic features
4. Implementation of speech processing algorithms
5. Implementation of musical features
   (music perception, MER and linguistics)
6. Simulations $\Rightarrow$ Musical features can help to improve emotion recognition in speech

Conclusions

Further research

- Environment: natural emotional speech, other languages
- Pattern Recognition steps: feature transformation, pitch extraction, classification...
- Improvement of musical features
  Dissonance model, Perceptual model of intonation, Emotionally meaningful moments
- Systematization of feature extraction step

"Even monkeys express strong feelings in different tones – anger and impatience by low, – fear and pain by high notes."

Charles Darwin, Naturalist

Thank you!

Looking forward to your questions...